



Europäisches Patentamt  
Europ an Patent Office  
Office européen des brevets



Publication number: **0 687 980 A2**

12

## EUROPEAN PATENT APPLICATION

21 Application number: 95105480.8

51 Int. Cl.<sup>6</sup>: G06F 13/12, G06F 13/42

22 Date of filing: 12.04.95

30 Priority: 17.06.94 US 261561

43 Date of publication of application:  
20.12.95 Bulletin 95/51

84 Designated Contracting States:  
DE FR GB

71 Applicant: International Business Machines  
Corporation  
Old Orchard Road  
Armonk, N.Y. 10504 (US)

72 Inventor: Capowski, Robert Stanley  
Milewood Road,  
RD 2, Box 49  
Verbank,  
New York 12585 (US)  
Inventor: Casper, Daniel Francis  
13, Brett Place  
Poughkeepsie,

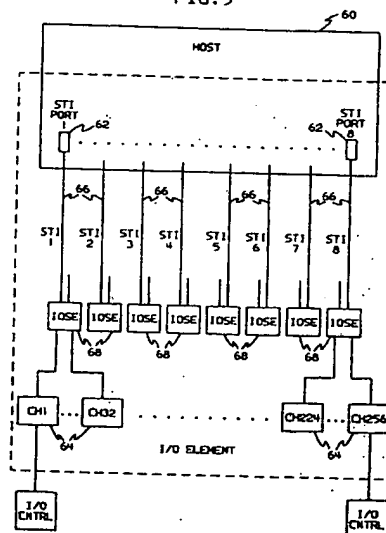
New York 12603 (US)  
Inventor: Cox, Frederick John  
6 Brook Hollow Court  
Wappingers Falls,  
New York 12590 (US)  
Inventor: Ferralolo, Frank David  
223 Spruce Street  
New Windsor,  
New York 12553 (US)  
Inventor: Halma, Marten Jan  
RR2, Box 24A,  
Hillside Road  
Poughquag,  
New York 12570 (US)

74 Representative: Schäfer, Wolfgang, Dipl.-Ing.  
IBM Deutschland  
Informationssysteme GmbH  
Patentwesen und Urheberrecht  
D-70548 Stuttgart (DE)

54 Enhanced input-output element

57 An enhanced input-output element has a self-timed interface (STI) in which a clock signal clocks bit serial data onto a parallel, electrically conductive bus and the clock signal is transmitted on a separate line of the bus. The received data on each line of the bus is individually phase aligned with the clock signal to provide a cost effective, modular, broadband, input/output element that can serve economically two channels and is modularly scalable to serve several hundred channels.

FIG. 5



EP 0 687 980 A2

This invention relates to an improved method and apparatus for transmitting digital data at high speeds via a parallel data bus, and more particularly, to a method and apparatus that provides a cost effective modular, high banded input/output element that can serve economically a few channels and is modularly scalable to serve several hundred channels.

The present United States patent application is related to the following co-pending United States patent applications incorporated herein by reference:

Application Serial No. 262,081, filed 06/17/94 (attorney Docket No. PO9-93-053), entitled "Digital Phase Locked Loop with Improved Edge Detector," and assigned to the assignee of this application.

Application Serial No. 261,515, filed 06/17/94 (attorney Docket No. PO9-93-054), entitled "Self-Timed Interface," and assigned to the assignee of this application.

Application Serial No. 261,522, filed 06/17/94 (attorney Docket No. PO9-93-056), entitled "Multiple Processor Link," and assigned to the assignee of this application.

Application Serial No. 261,603, filed 06/17/94 (attorney Docket No. PO9-93-058), entitled "Massively Parallel System," and assigned to the assignee of this application.

Application Serial No. 261,523, filed 06/17/94 (attorney Docket No. PO9-93-059), entitled "Attached Storage Media Link," and assigned to the assignee of this application.

Application Serial No. 261,641, filed 06/17/94 (attorney Docket No. PO9-93-060), entitled "Shared Channel Subsystem," and assigned to the assignee of this application.

As will be appreciated by those skilled in the art, such factors as noise and loading limit the useful length of parallel busses operating at high data rates. In the prior art, the length of the bus must be taken into account in the system design and the bus length must be precisely as specified. Manufacturing tolerances associated with physical communication link (chips, cables, cord wiring, connectors, etc.) and temperature and variations in power supply voltage also limit the data rates on prior art busses comprised of parallel conductors. Further, many prior art computer systems transfer data synchronously with respect to a processor clock, so that a change in processor clock rate may require a redesign of the data transfer bus.

General purpose computer systems are comprised of central processing element(s), local memory (storage) elements and input/output (I/O) elements. Of these, the processing and storage elements have been fitted directly from the continuous advances being made in circuit density, enabling more and more processing capability/storage per

chip.

The I/O element, gated by other physical constraints, has not been able to track the performance gains of the processing/storage elements. Hence, the I/O element can limit the overall computer systems' performance in many applications. Further aggravating this problem are new system applications, such as client/server, that require large numbers of I/O, and multimedia that demand higher performance I/O elements. It also follows that for many applications, the smaller and smaller, yet faster, central processor elements (benefitting from circuit technology density advances) require more and more I/O capability to satisfy the increased demand for data (to be moved and/or to be processed). These trends can be expected to persist. Given that a primary role of a computer system I/O element is that of a transformer of bus speed (as well as of protocol and technology) between increasingly faster processor/storage elements and the slower, relatively unchanging speed of the attached I/O controllers, plus increases in required connectivity will continue to drive the need for improved efficiency and performance of the I/O element.

The I/O element often takes the form, especially on the most powerful systems, of a large multiplexer feeding and being fed by a network of more numerous and slower buses/links (becoming more numerous and slower the further away from the host they are in the network).

A current typical large computer system I/O element can require more than 100 channel functions as the middle stage/level in a hierarchically arranged busing network (between the highest level internal bus (fastest) and the more numerous and slower I/O controllers). The channel functions are packaged as part of the computer system and provide the interfaces to communicate, over distance (km), with the I/O controllers.

The large number of channel functions required, the channel transceiver design, the power required for distance and the connectors needed for external cable attachment all impede I/O element miniaturization.

In the prior art system, efforts are made in the I/O element design to avoid requiring excessive host signal pins (obtainable only through exotic/expensive packaging). Toward that end, minimum width data path/bus connections are employed between the channels and higher levels (internal buses) of the network. Even so, the number of signal pins, internal cables, connectors and circuit board complexity remains high.

The narrow buses are cycled faster than wider buses to achieve equivalent data transmission performance; however, this restricts maximum bus (cable) length. This trade-off between cable dis-

tance/net length and cycle time is typical with most prior art bus/link technologies and is a significant packaging constraint. Optimal packaging for cost and extendibility is compromised by requiring all channels to be within a limited (e.g., 3 meters) cable distance from the processor storage/control elements. This is especially troublesome in the more powerful systems where the system package must accommodate 100 or more channels.

Then too, prior art systems operate the numerous channels/buses synchronously to the host system clock to avoid inflating host circuit count via additional synchronizing buffers and logic. The channels/buses operating cycle time ranges are slower than the host, thus clock multiplier logic is used to provide clocks that are a multiple of the host clock cycle. This adds complexity to the design (as different multipliers can be required when host cycle times change) to remain within channel/bus operating ranges.

An object of this invention is the provision of a cost effective bus data transfer system that can operate at high data transfer rates without tight control of the bus length, and without system clock constraints; a system in which the maximum bus length is limited only by the attenuation loss in the bus.

Another object of the invention is the provision of a general purpose, low cost, high performance, point to point data communication link where the width and speed of the interface can easily be modified to tailor it to specific bandwidth requirements and to specific implementation technologies, including VLSI technologies.

A further object of the invention is the provision of a bus data transfer system that operates a clock speed equal to the data rate.

A more specific object of the invention is the provision of a system that adjusts the phase or arrival time of the incoming data on the receive side so it can be optimally sampled by the local receive clock, compensating for many of the manufacturing tolerances associated with the physical link (chip, cable, card wiring, connectors, etc.) as well as temperature changes and power supply output variations.

Another object of the invention is the provision of an enhanced I/O element to facilitate decentralization of the I/O element into (n) sub-elements, with significant advantage to system cost, flexibility and extendibility.

One more object is to allow a customized array of different channel adaptors to be implemented within the I/O element without impacting expensive specific hardware and software.

A still further object of the invention is to enable each I/O sub-element to operate asynchronously to the host clock, thereby avoiding the clock

control complexity associated with the previously described variable clock multiplier logic. This permits the I/O sub-elements to be used by a wider range of hosts/hostcycle times.

Briefly, this invention contemplates the provision of a self-timed interface (STI) in which a clock signal clocks bit serial data onto a parallel, electrically conductive bus and the clock signal is transmitted on a separate line of the bus. The received data on each line of the bus is individually phase aligned with the clock signal. The received clock signal is used to define boundary edges of a data bit cell. It is defined individually for each line and the data on each line is individually phase adjusted so that, for example, a data transition position is in the center of the defined cell. The data is read into a buffer where it is decoded and read out synchronously with the receiver system clock. At the data rates contemplated in the application of this invention, the propagation delay is significant. However, within limits, the buslength is not critical and is independent of the transmit and received system clock. The phase adjustment can compensate for a skew of up to one bit cell across the width of the bus.

In one specified embodiment of the invention, data to be transmitted is transferred to a buffer synchronously with the transmitter system clock, which may or may not be the receiver system clock. A controller formats the data into packets for byte parallel, bit serial, transmission along with headers specifically coded to provide unique data patterns that allow for correction of skew of up to three bit cells in addition to the initial phase adjustment.

The foregoing and other objects, aspects and advantages will be better understood from the following detailed description of a preferred embodiment of the invention with reference to the drawings, in which:

Figure 1 is an overview block diagram illustrating the application of a self-timed interface, in accordance with the teachings of this invention, to data communication among computer chips.

Figure 2 is a block diagram illustrating one embodiment of a transmitter serializer for implementing a self-timed interface in accordance with this invention.

Figure 3 is a block diagram illustrating byte synchronization in accordance with the invention.

Figure 4 illustrates phase alignment and sampling logic in accordance with a preferred embodiment of the invention.

Figure 5 illustrates an input/output element configuration in accordance with the teachings of this invention.

Referring now to Figure 1 of the drawings, it illustrates one embodiment in which a self-timed interface in accordance with the teachings of this invention can be used. This exemplary embodiment of the self-timed interface provides data communications between two microprocessor chips, labeled here as Chip A and Chip B. However, as will be apparent to those skilled in the art, the self-timed interface of this invention is applicable to provide data transfer between a wide variety of components or nodes.

Chip A has a transmit port labeled 12A and Chip B has a transmit port labeled 12B. Similarly, Chips A and B have receive ports labeled 14A and 14B, respectively. The ports are connected by two self-timed interface busses 16; one for each transmission direction. In this exemplary embodiment of the invention, each bus 16 is one byte wide, and comprised of nine electrical conductors; eight conductors for data and one conductor for a clock signal.

Each transmit port (12A and 12B) includes a transmit logical macro 18 that provides a logical interface between the host logic and the self-timed interface link 16. Sync buffers 22 provide an interface between the host clock and the self-timed interface clock. This allows the self-timed interface link to run at a predetermined cycle time that is independent of the host clock, making the self-timed interface link independent of the host. An outbound physical macro 24 serializes a word-wide data flow into a byte-wide data flow that is transmitted along with the clock on the self-timed interface link 16.

Each receive port (i.e., 14A and 14B) includes an inbound physical macro 26 that first dynamically aligns each data bit with the self-timed interface clock signal. It aligns any bits with skew up to three bit cells and deserializes the bytes into words. A receive logical macro 28 provides an interface between the self-timed interface receiver logic and the host logic and generates link acknowledge signals and link reject signals, which are coupled by internal links 33 and transmitted back to the transmitting port via an outbound self-timed interface link 16. In order to compensate for variations in electrical path delay, the phase of the incoming data is adjusted, or self-timed. Each bit (line) is individually phase aligned to the transmitted reference clock and further aligned to compensate, within embodiment, for up to three bit cells of skew between any two data lines. The self-timing operation has three parts. The first is to acquire bit synchronization; the second is byte/word alignment; and the third is maintaining synchronization.

In acquiring bit synchronization, the link takes itself from a completely untimed state into synchronous operation. Any previous condition on the STI

interface or logic is disregarded with a complete logic reset. The bit synchronization process can be rapidly established, for example on the order of 200 microseconds. The phase of the incoming data is manipulated on a per line basis until the data valid window or bit interval is located. This is accomplished using a phase detector that locates an average edge position on the incoming data relative to the local clock. Using two phase detectors one can locate two consecutive edges on data and these two consecutive edges define the bit interval or data valid window. The data to be sampled by the local clock is the phase of the data located halfway between the two edges of the data.

Byte alignment takes place by manipulating the serial data stream in whole bit times to properly adjust the byte position relative to a deserializer output. Word alignment takes place next by manipulating the deserializer data four bit intervals a time to ensure proper word synchronization on the STI interface. A timing sequence allows proper bit, byte and word synchronization.

Synchronization maintenance occurs as part of the link operation in response to temperature and power supply variations.

Referring now to Figure 2, which illustrates one embodiment of a transmit serializer for a bit serial byte parallel interface used in the practice of the invention. Here a four byte wide data register 23 receives parallel inputs 25 (bytes 0, 1, 2 and 3 inputs shown here) and multiplexes them to a one byte wide self-timed interface. Data is clocked from the register 23 by a divide by two of the self-timed interface clock on line 27. Bit zero from bytes 0, 1, 2 and 3 are serialized and transmitted on link 0 of the self-timed interface, shown here. Bit 1 from bytes 0, 1, 2 and 3 will be transmitted on link 1 (not shown) and so on.

To minimize the bandwidth requirements of the communication media the STI clock is one half the frequency of the transmitted data (baud) rate, i.e., a 75 Mhz clock will be used for a 150 Mbit/S data rate. The clock will be generated from an STI oscillator source, this is done to decouple the system or host clock from the STI link. The data will be transmitted with both edges of the clock.

Referring now to Figure 3, assuming a bit synchronization process as described in connection with co-pending United States patent application, Application Serial No. 261,515, filed 06/17/94 (attorney Docket No. PO9-93-054), entitled "Self-Timed Interface," and assigned to the assignee of this application, byte synchronization starts by coupling the phase aligned data (now 2 bits wide) into a shift register 33 whose outputs are coupled to multiplexer 35. The control inputs to the multiplexer will be used to deskew the particular data line from the other data lines by whole bit times. The de-

serializer data output for a particular data line is monitored for the expected timing pattern (X 0 1 0 where X is a don't care) to determine the proper order of the received data. If at any time a zero is detected in the bit 3 position the multiplexer is incremented thus moving the byte boundary by one bit time. This process is repeated until the proper byte boundary is located. The multiplexer control wraps around from a binary 3 to a binary 0 in case the correct position was incorrectly passed through the previous time. This function allows synchronization of data lines skewed by more than an entire bit time.

Referring now to Figure 4, in this embodiment of the invention, the clock rate is the same as the data rate. The data edges that define a data window are each detected independently of the other and the data is sampled at the midpoint between the edges when the edges have been aligned with the clock. The position of the edges of incrementally separated phases of the input data stream are successively compared to the position of the rising and falling edges of the clock in order to locate the edges of the data stream with respect to both edges of the clock (e.g., the rising and falling edges).

The data phase pairs are generated in this specific embodiment of the invention by three incrementally selectable delay elements 80, 82, and 84. For example, the elements 80 and 82 provide delays, respectively, in 1/10th and 1/5th bit time increments and element 84 provides fine increments on the order of 1/20th of a bit time. The fine delay element 84 is separated into three groups to provide early edge detection, system data detection, and late edge detection. An early guard band selector 86 successively selects one phase of the data stream to provide an "early" phase of the incrementally separated phases - one for the rising edge and one for the falling edge. Similarly, a late guard band selector 90 successively selects one phase of the data stream to provide a "late" phase of the incremental phases - again one for the rising edge and one for the falling edge. A selector 88 selects incremental phases for the mid-cell system data position.

A selected data phase is coupled as an input to master-slave RES-FES latch pairs 92, 94, and 96. The rising edge data samples are clocked into the RES latches and the falling edge data samples are clocked into the FES latches. The outputs of the RES-FES latch pair 92 are connected to an early edge detector 98. Similarly, the outputs of the RES-FES latch pair 96 are coupled to a late edge detector 100. The RES latch of pair 94 is coupled to the early edge detector 98 and the FES latch of pair 94 is coupled to the late edge detector 100. Each edge detector (98 and 100) outputs a "lead",

a "lag" or a "do nothing" output which indicates the location of a data edge with respect to the reference clock edge location. The output of each edge detector is coupled via a suitable filter 102 (i.e., a random walk filter), back to its respective selector 86 and selector 90, respectively. The selectors shift the phase of the data coupled to the RES-FES latches in the direction indicated, or if "do nothing" is indicated, the phase of the data at that edge is not shifted.

Data control logic 104 controls the system data output by selecting the phase of the data that is halfway between the two data edges when the data edges are aligned with the reference clock. A phase of the data (Data 1 and Data 2) is outputted at each reference clock edge.

In operation of a specific embodiment, at power on the logic will automatically begin the bit synchronization process. A 16 microsecond (final value TBD) timer is started, the bulk delays are reset to their minimum delay and a 16 bit counter running off the divided down clock is started. The edge detect circuitry will sample the incoming data with the received reference clock. The edge detector will output a "lead", a "lag" or a "do nothing" signal that indicates the data edge location relative to the reference clock. This signal is filtered by a Random Walk Filter (RWF) and fed-back to the selectors of their respective RES and FES circuits. The selectors shift the phase of the data into the RES and FES as indicated by the edge detector. Each edge detector operates independently of the other. Each will locate the transitions on data relative to the received (ref) clock by manipulating the incoming phase of the data into the edge detector as described above. The phase of the system data is controlled by the data control logic which selects the phase of the data halfway between the two edge detectors. In parallel with the bit synchronization process, the order of bits out of the deserializer are manipulated to the correct order (see byte/word synchronization below). When the 16 microsecond timer trips the algorithm resets a deserializer error latch and restarts the 16 microsecond counter. The deserializer output is compared against the expected timing pattern (X 0 1 0 where X is a don't care). A single miscompare on any cycle during the next 16 microseconds will set the deserializer error latch. When the 16 microsecond counter trips again the algorithm checks the addresses of the EGB, LGB, and data selectors, deserializer error latch. In order for a bit to end the initial bit synchronization search state, the deserializer output latch must have remained set AND the all selectors must be properly centered in their tracking range (centering ensures that adjustments can be made to allow for the tracking of temp. and power supply variations after the initial

bit synchronization process). If both conditions are not met then the algorithm adds a bulk delay element, resets the 16 microsecond counter and the search process begins once again. Each and every bit (data line) on the STI interface undergoes this process in parallel. Once an individual data line is determined to meet the initial bit synchronization criteria described above it is degated while the other lines continue to be adjusted. The bit synchronization process is complete once all bits are adjusted and meet the search criteria. The logic will not exit the bit synchronization mode until the 16 bit counter trips.

Finally word alignment takes place. Word alignment is established by manipulating the deserializer output bus four bits at a time until word synchronization is established. Note that the first register is shifted by four bit times relative to the second register. Four bit times is the maximum any data bit can be skewed relative to another data bit (3 bit times on link + 1 bit time from phase alignment section).

During normal operation the physical macro will continuously monitor the incoming data to ensure that the optimum clock sampling relationship exists. Small updates will be made to track temperature, power supply and data jitter. These updates will be seamless and transparent to the host logic. Approximately 1/2 a bit time of delay will be needed to compensate for temperature and power supply variations to maintain proper synchronization. This added delay is in the fine delay elements section. There is also circuitry to monitor the position of the guard bands relative to the allowable range of operation. If a guard band reaches the end of its range, two cases exists: 1) a new bulk delay element is added and the fine delay elements are adjusted accordingly. Note this can cause sampling errors in the data. The circuitry that makes these on the fly bulk adjustments can be inhibited so no on the fly bulk delay adjustments are made during normal operation. The second case exists when one of the guard bands reaches the end of its range and the on the fly bulk delay adjustment is inhibited, the physical macro will signal the logical STI macro that a bit synchronization is required soon. The link should finish the immediate work and force the link into timing mode.

Referring now to Figure 5, a host processor 60 in this exemplary embodiment needs only eight STI ports 62 to support up to 256 input-output channels 65, which are in turn coupled to input-output controllers 65. STI links 66 couple the ports 62 to input-output sub-elements 68 labeled IOSE. The sub-elements 68, which may be integral with or physically located away from the host 60, each can support up to 32 channel functions. The extended operating length of the self-timed interface

66 (i.e., up to ten meters without performance degradation) makes it possible to package this many channel functions in a cost effective manner.

A single STI is capable of supporting 32 channel functions. For those system configurations requiring only up to 32 channels therefore, only one STI port, cable and I/O sub-element need be used. This enables incremental cost advantages over current centralized I/O elements in which even the smallest channel configuration can incur the cost of hardware in place for the maximum (e.g., 256) channel configuration.

STI enables each I/O sub-element to operate asynchronously to the host clock, thereby avoiding the clock control complexity associated with the previously described variable clock multiplier logic. This permits the I/O sub-elements to be used by a wider range of hosts/host cycle times. Because of the high performance of STI, fewer are needed, thus the re-synchronizing buffers/logic at each end is not burdensome.

While the invention has been described in terms of a single preferred embodiment, those skilled in the art will recognize that the invention can be practiced with modification within the spirit and scope of the appended claims.

#### Claims

1. An input-output element for coupling digital data at high data rates between a host processor and a plurality of input-output controllers, comprising in combination:
  - a host processor;
  - a high-speed, self-timed interface;
  - an input-output channel sub-element;
  - a plurality of input-output channels connecting said input-output sub-element respectively to a plurality of input-output controllers;
  - said high-speed, self-timed interface connecting said host processor to said input-output channel sub-element;
  - said high speed, self-timed interface including a transmitting node for transmitting a digital data and a clock signal and a receiving node for receiving said digital data and said clock signal, said transmitting node connected to said receiving node by a parallel data bus to individual lines of which a digital data stream is coupled by said clock signal at said transmitting node,
  - said bus including a separate line for transmitting said clock signal, and said receiving node including means to phase align a digital data stream on each of said lines separately with respect to said clock signal.

2. An input-output element as in claim 1 wherein said input-output sub-element and said self-timed interface each support up to 32 channels.
3. An input-output element as in claim 1 wherein each input-output sub-element operates asynchronously with respect to a host clock signal.

5

10

15

20

25

30

35

40

45

50

55

FIG. 1

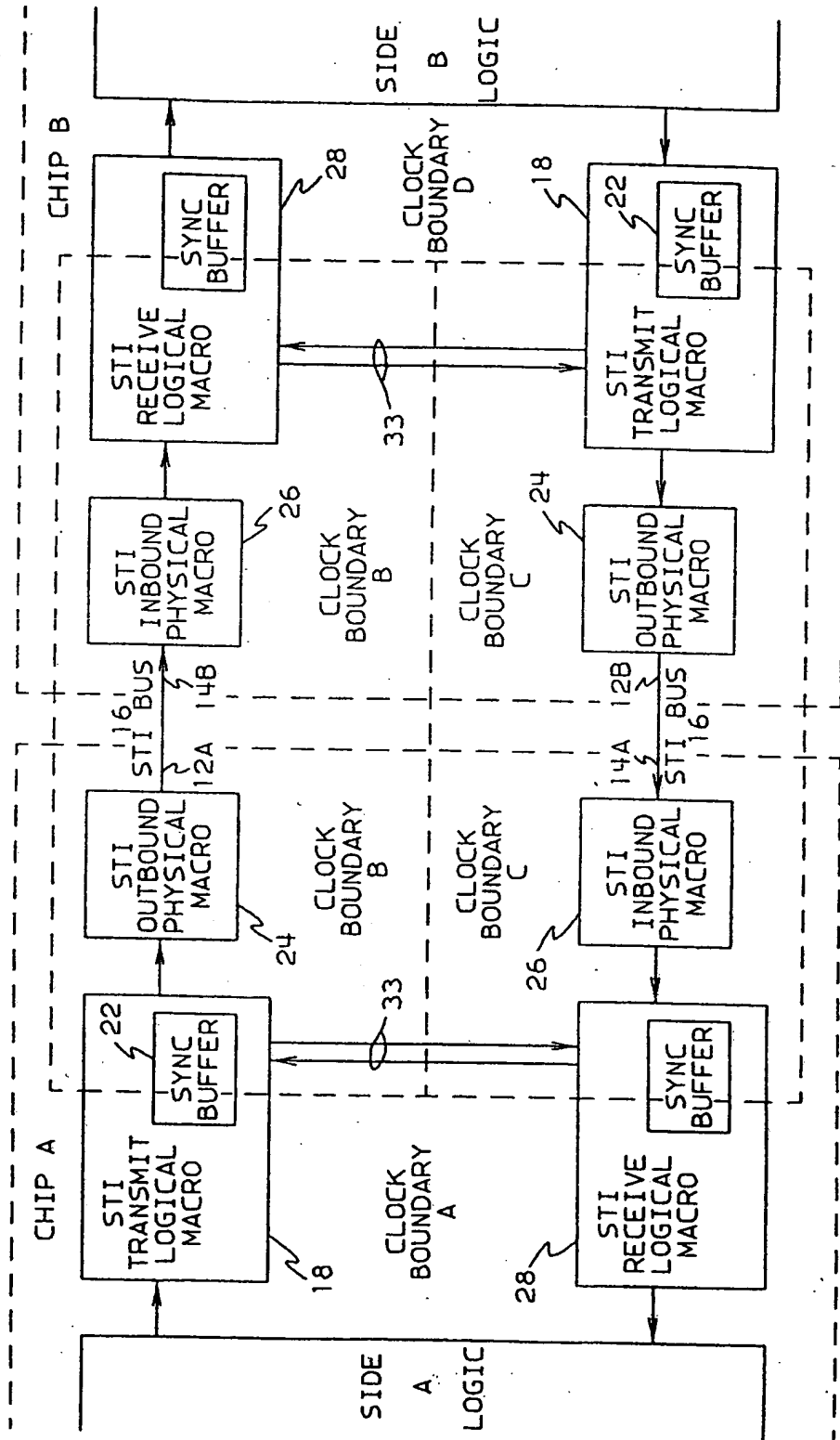




FIG.2

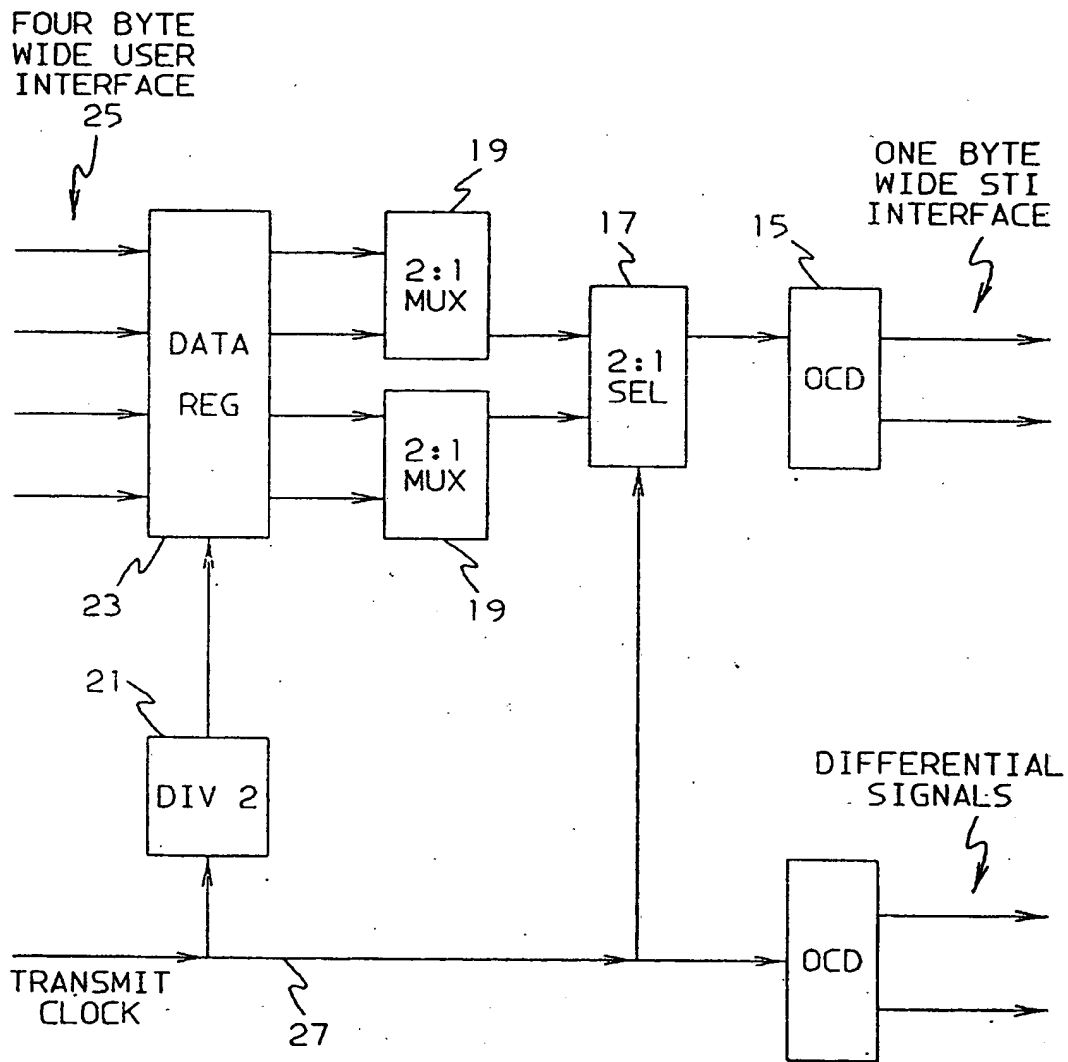


FIG. 3

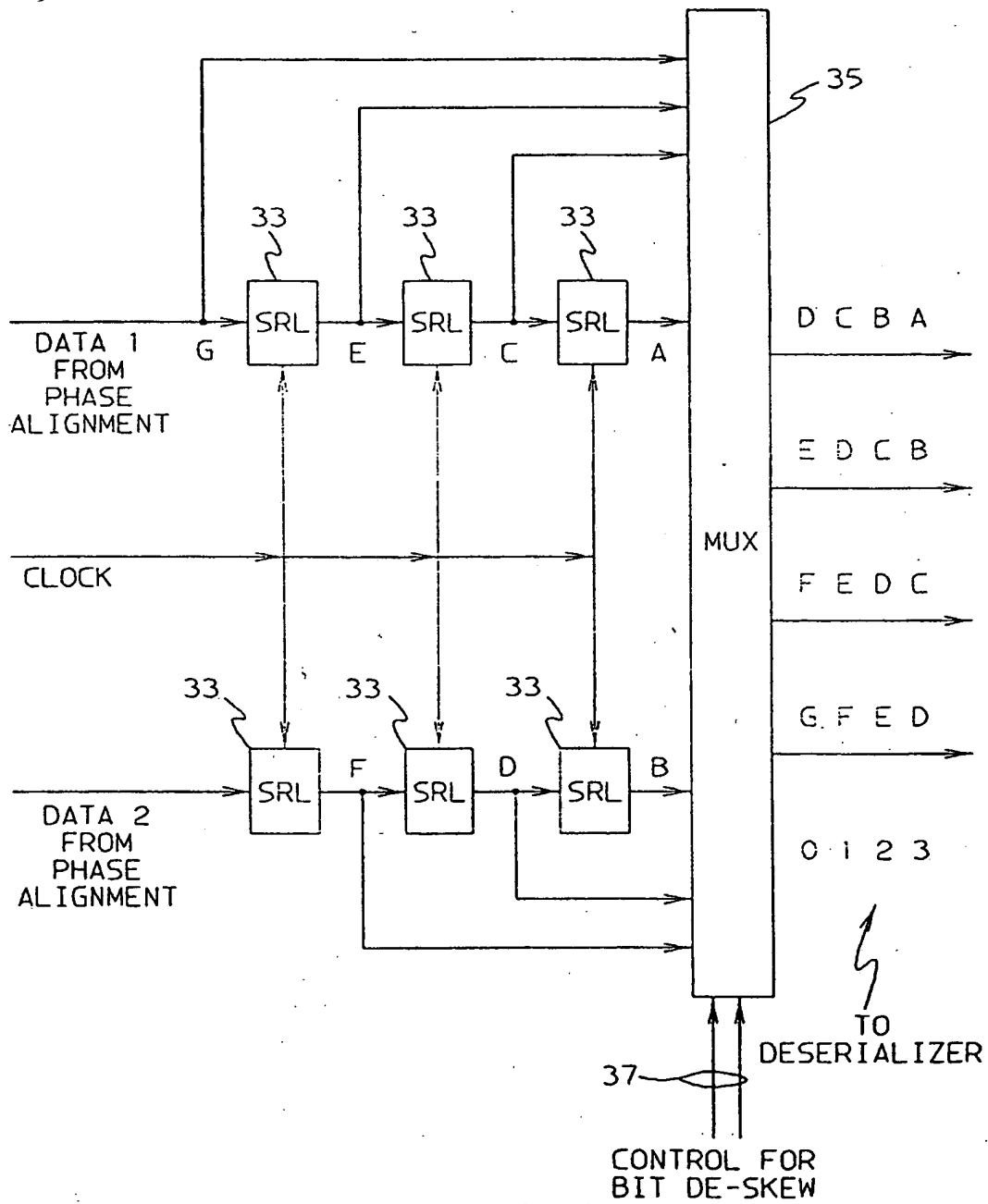


FIG. 4

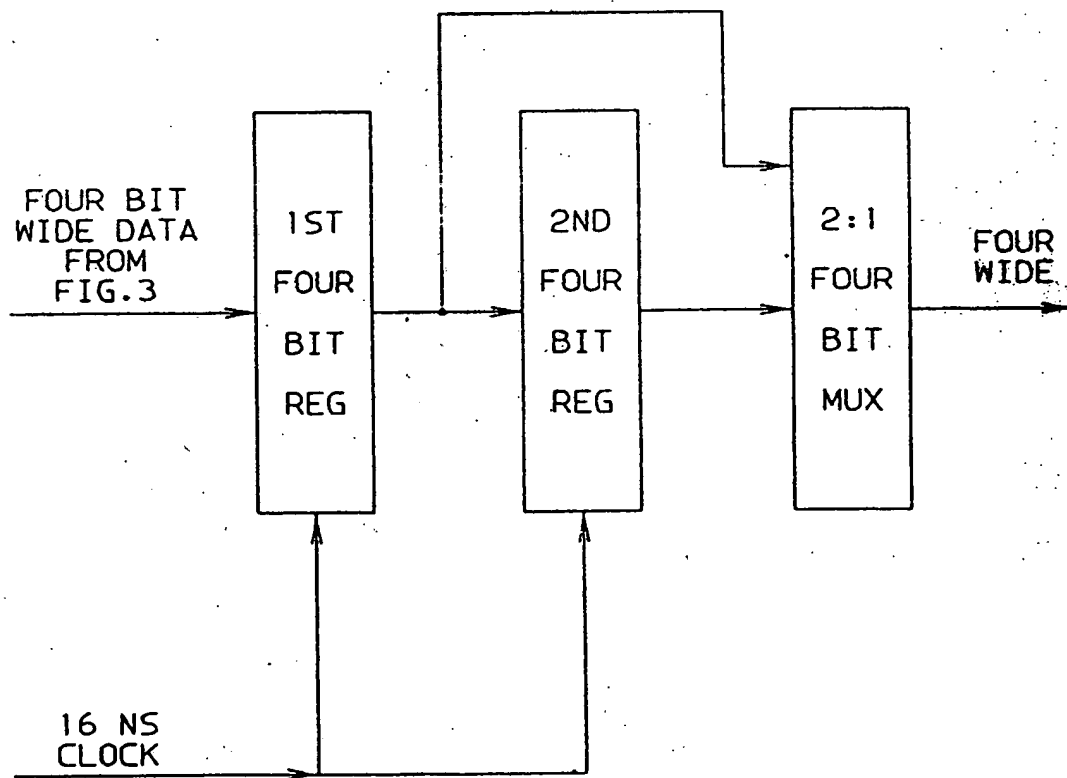


FIG. 5

